Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done?

Mary Lou Thompson, J E Myers, D Kriebel

Abstract

***Objectives*—To review the appropriateness of the prevalence odds ratio (POR) and the prevalence ratio (PR) as effect measures in the analysis of cross sectional data and to evaluate different models for the multivariate estimation of the PR.**

***Methods*—A system of linear differential equations corresponding to a dynamic model of a cohort with a chronic disease was developed. At any point in time, a cross sectional analysis of the people then in the cohort provided a prevalence based measure of the effect of exposure on disease. This formed the basis for exploring the relations between the POR, the PR, and the incidence rate ratio (IRR). Examples illustrate relations for various IRRs, prevalences, and differential exodus rates. Multivariate point and interval estimation of the PR by logistic regression is illustrated and compared with the results from proportional hazards regression (PH) and generalised linear modelling (GLM).**

***Results*—The POR is difficult to interpret without making restrictive assumptions and the POR and PR may lead to different conclusions with regard to confounding and effect modification. The PR is always conservative relative to the IRR and, if PR>1, the POR is always >PR. In a fixed cohort and with an adverse exposure, the POR is always ⩾IRR, but in a dynamic cohort with sufficient underlying follow up the POR may overestimate or underestimate the IRR, depending on the duration of follow up. Logistic regression models provide point and interval estimates of the PR (and POR) but may be intractable in the presence of many covariates. Proportional hazards and generalised linear models provide statistical methods directed specifically at the PR, but the interval estimation in the case of PH is conservative and the GLM procedure may require constrained estimation.**

***Conclusions*—The PR is conservative, consistent, and interpretable relative to the IRR and should be used in preference to the POR. Multivariate estimation of the PR should be executed by means of generalised linear models or, conservatively, by proportional hazards regression.**

(*Occup Environ Med* 1998;**55**:272–277)

Keywords: prevalence; cross sectional study; logistic regression

Department of Biostatistics, Box 357232, University of Washington, Seattle, WA 98195, USA
M L Thompson

Department of Community Health, Medical School, University of Cape Town, Rondebosch, South Africa
M L Thompson
J E Myers

Department of Work Environment, University of Massachusetts, Lowell, MA 01854, USA
D Kriebel

Correspondence to:
Dr M L Thompson,
Department of Biostatistics,
Box 357232, University of
Washington, Seattle, WA
98195, USA.

Accepted 31 October 1997

In the analysis of data from cross sectional studies, two ratio measures of effect suggest themselves: the prevalence odds ratio (POR) and the prevalence ratio (PR), sometimes incorrectly called the prevalence rate ratio. The choice between these two has been the source of ongoing debate in the epidemiological literature over the past few years.[1–16] Although there is no dispute that the PR and POR will be similar for a rare disease, they may be very discrepant for a common disease, and common diseases are often the focus of cross sectional studies. A recent paper has illustrated the divergence of the POR and PR for different underlying disease prevalences.[17]

The debate acknowledges these differences and has had two main thrusts: firstly, discussion of which of the two effect measures is the more appropriate; and secondly, disagreement on the appropriate model with which to construct multivariate estimates of the PR and its standard error.

We present an analysis which clarifies the relations between PR, POR, and the incidence rate ratio (IRR) in cross sectional studies of chronic disease, and use this to make recommendations about the appropriate ratio measure of effect. We also derive an expression for the variance of the log of the estimated PR from logistic regression which involves standard results for the variance and covariance of the logistic regression coefficients, thus enabling use of widely available logistic regression packages to carry out PR analyses.

Measures of effect in cross sectional studies

Cross sectional studies are conducted with many objectives; sometimes the interest is primarily descriptive, but increasingly these studies are used, despite their well known limitations, to seek aetiological information. Interest may be in drawing inferential conclusions from cross sectional studies of prevalent conditions because of cost, or the difficulty of the collection of incidence data. For example, in developing countries where public health and demographic data bases may be scarce, there may be little choice but to work with prevalence surveys, at least in preliminary studies. In occupational studies of subjective conditions—such as respiratory symptoms—there are rarely alternatives to cross sectional studies. But these studies are not necessarily seen as purely descriptive; progressively more publications suggest that respiratory symptoms can be studied quickly and cheaply as early markers of chronic conditions. This may be useful in identifying hazardous exposures instead of waiting for the eventual development of chronic disease. Despite the serious limitations of cross sectional studies it is important to understand their inferential capabilities, and in particular the relations between measures of association from these studies and the preferred measures. These are the IRR (if studying a dynamic population) or the cumulative incidence ratio (CIR, if studying a closed cohort). We will not discuss difference measures of effect in this paper.

One reason often cited for studying the POR is that, under certain assumptions called stationarity, and if the duration of disease in the exposed and unexposed groups is equal, the POR will estimate the IRR, whereas the PR will not.[6 18 19] The assumption of equal duration of disease is questionable in many practical settings. If one is studying a symptomatic condition in which subjects may act to avoid illness by removing themselves from exposure, then the duration of disease may well be substantially shorter among the exposed than among the unexposed population. Occupational cross sectional studies of acute conditions like respiratory or musculoskeletal disorders often find that a short duration of exposure (employment) seems to be a risk factor for the condition[20]—that is, long duration of exposure occurs primarily among those who are less susceptible to the disease, an example of the healthy worker survivor effect. This is a concrete and all too common example of a violation of the assumption of equal disease duration under which POR=IRR.

It must also be born in mind that stationarity applies to an acute condition which resolves completely after a short duration of disease in the context of a fixed cohort. It is doubtful how often, if at all, these rather restrictive stationarity assumptions can be met, particularly in occupational epidemiological studies.

Almost all participants in this debate stress the need for an effect measure to have natural intelligibility. Lee and Chia[3] state that "Whereas PR is easy to interpret and to communicate, POR lacks intelligibility" and Axelson *et al*[7] state that "The resulting prevalence odds ratio may be ... without any clear interpretation in terms of risk". Once one leaves the stationary setting, the interpretation of the POR becomes unclear, or worse still, it is incorrectly implicitly interpreted as a risk ratio. Use of the PR in cross sectional studies ensures "truth in advertising": everyone knows that prevalence rather than incidence is an inferior basis for measurement in causal inference. Hence labelling a measure of effect a PR is to declare it of limited inferential value. Lacking longitudinal data the best one can do is to estimate a PR and let the reader beware.

In certain specialised settings, the POR may have intuitive meaning.[10] In ecological designs, one may choose as the object of study, various ratios—such as the sex ratio or the prevalence of smoking—and compare these among groups hypothesised to differ on these measures. It would then be appropriate to use POR as a measure of effect. But in this paper, we focus on the more common situation in which one seeks to estimate a relative measure of disease occurrence in two or more groups.

It has also been noted that the POR and PR may behave differently with regard to patterns of confounding. The published debate[7 11 12] includes a discussion of the possibility that the choice of effect measure will influence whether a covariate is identified as a confounder or effect modifier (or neither). In the light of these arguments against the use of the POR, the fact that it may be associated with different patterns of confounding or effect modification than the arguably preferable measure, the PR, is of concern. This may result in bias in the form of analysis deviation.[21] This issue is of interest beyond the debate around choice of effect measure. It raises questions as to the definition of confounders and effect modifiers[22] and their aetiological interpretation. It is additionally possible for the choice of different statistical models to induce different patterns of confounding or effect modification. We will allude to this again later, but the broader arguments around this issue are peripheral to the focus of the present paper.

Relations between IRR, PR, and POR

Most of the conditions studied in occupational epidemiology are chronic—that is, of slow onset, long duration, and irreversible. Aetiological research with cross sectional study designs is typical in less developed settings. This is due to the absence of databases supporting cohort or follow up studies and other more complex designs. In this context it is of interest to determine how two estimators of effect based on cross sectional data, the PR and POR, perform relative to the underlying effect as measured by the incidence rate ratio as being the most general measure of effect.

To explore the relations between the IRR, PR, and POR that would be expected in a study of a dynamic occupational cohort with a chronic disease, we considered a setting in which initially disease free exposed and unexposed workers are followed up over time for the development of disease. Workers who leave the cohort are assumed to be replaced by disease free workers, regardless of their disease status on departure. At any fixed point in time, one may then consider a cross sectional analysis of those currently in the cohort and how it relates to the follow up of the dynamic cohort which would provide the ideal measure of exposure effect.

The scenario is thus a dynamic cohort of exposed, unexposed, diseased and disease free people, which it is possible to represent as a deterministic system of linear differential equations. By solving these equations, we are able to make general inferences about the relations between the three effect measures IRR, PR, POR in this setting. We acknowledge the simplicity of the setting. We are, for instance, considering only a single dichotomous risk factor and assuming that disease and exodus rates remain constant over time (or with age) and we are not considering stochastic variation. However, the setting is nevertheless realistic enough that the results that we are able to show (by

Table 1 The association between IRR, PR, and POR

Follow up time (y)		*5*	*10*	*15*	*20*	*50*
IRR=5:						
	PR	4.8	4.6	4.5	4.3	3.5
p_1=0.015, p_2=0.003	POR	5.1	5.2	5.4	5.4	5.7
α_1=0.005, α_2=0.001	Prevalence (%)	4	8	12	15	31
	PR	4.5	4.0	3.7	3.4	2.3
p_1=0.05, p_2=0.01	POR	5.4	5.9	6.3	6.7	7.0
α_1=0.01, α_2=0.005	Prevalence	13	23	31	38	57
	PR	3.1	2.2	1.7	1.4	0.9
p_1=0.15, p_2=0.03	POR	4.8	3.7	2.8	2.1	0.8
α_1=0.1, α_2=0.01	Prevalence	28	40	46	50	62
IRR=1.5:						
	PR	1.5	1.5	1.4	1.4	1.3
p_1=0.015, p_2=0.01	POR	1.5	1.5	1.5	1.6	1.6
α_1=0.002, α_2=0.001	Prevalence	6	11	17	22	44
	PR	1.4	1.4	1.4	1.3	1.1
p_1=0.03, p_2=0.02	POR	1.5	1.5	1.5	1.5	1.4
α_1=0.01, α_2=0.005	Prevalence	11	21	29	36	61
	PR	1.3	1.1	1.0	0.9	0.8
p_1=0.1, p_2=0.07	POR	1.4	1.2	1.0	0.8	0.4
α_1=0.05, α_2=0.01	Prevalence	31	49	60	66	76

virtue of its tractability) provide broader insight. In our interpretation of the results we will generally assume that IRR>1, so we are looking at an adverse rather than a protective effect.

Consider then a cohort which starts with n exposed and m unexposed people, all of whom are disease free at the start of follow up and where p_1 is the disease incidence rate for the exposed people and p_2 is the equivalent rate in the unexposed people. Assume further that exposed people leave the cohort at the rate of α_1 and unexposed at the rate of α_2. Those who leave the cohort (whether diseased or non-diseased) are replaced by disease free people.

The IRR remains constant at $p_1/\ p_2$, but the PR and POR vary according to duration of follow up. Appendix 1 shows that this allows the determination of the following associations between the effect measures. If IRR>1 and $\alpha_1+p_1>\alpha_2+p_2$ (the overall rate of exodus from the disease free exposed—whether by acquisition of the disease or by departure from the workforce—is greater than the corresponding exodus from the unexposed), it can be shown that PR_t<IRR for all t. Similarly, if IRR<1 and $\alpha_1+p_1<\alpha_2+p_2$, then PR will be greater than IRR. In general, (provided the exodus rates are not in the opposite direction to and greater than the disease rates), the PR is always conservative relative to the IRR (closer to the null effect). It is easily seen from the results in appendix 1 that, asymptotically, $PR\rightarrow IRR(\alpha_2+p_2)/(\alpha_1+p_1)$ as follow up time increases.

The development in appendix 1 also allows one to show that $POR_t>PR_t$, provided $PR_t>1$, which is consistent with several other authors.[4 17] Asymptotically, it can be seen that $POR\rightarrow PR\alpha_2(\alpha_1+p_1)/(\alpha_1(\alpha_2+p_2))$ as follow up time increases. Also, for a dynamic cohort, POR_t>IRR for short follow up (for small values of t) but, asymptotically, $POR_t\rightarrow IRR\ \alpha_2/\alpha_1$ as follow up time increases, and hence, for $\alpha_1>\alpha_2$, POR_t<IRR for large t.

When there is equal exodus from the exposed and unexposed groups ($\alpha_1=\alpha_2$), $POR_t\geqslant IRR$ and one has $POR_t\rightarrow IRR$ for large t. For a fixed cohort ($\alpha_1=\alpha_2=0$), $PR_t\rightarrow 1$, for large t. It should be noted that the effect of the relative magnitudes of α_1 and α_2 on these results also supplies a framework for considering aspects of the healthy worker effect.

Table 1 provides examples illustrating the consequences of these results for different incidence and exodus rates applied to generate cohorts with cross sectional data sets accruing after a given period. Rates are expressed as cases per person-year of observation. The prevalence (%) at each time cross section (in years) is provided in the last row, for the setting of equal numbers of exposed and unexposed workers at the start of the follow up. We again assume that we are looking at a chronic disease, and that all those leaving the cohort are replaced.

In summary, when the IRR>1, the PR<IRR and the POR gives a larger measure of effect than the PR (subject to the given conditions). In a fixed cohort, POR⩾IRR (provided p_1 > p_2). For a dynamic cohort it is possible to show that, with sufficient duration of follow up, POR will always eventually be less than the IRR (provided the exodus from the exposed is greater than the exodus from the unexposed). Thus the PR is conservative relative to the IRR, but the POR may underestimate or overestimate depending on the underlying effective duration of follow up and whether the exodus from the cohort is differential or not with respect to exposure. The duration of follow up referred to here derives from the incidence study generating the summary data for the cross sectional analysis. The fourfold table (a_t, b_t, d_t, e_t) at the end of any specified period is used for computing the PR and the POR.

The results of this analysis of the relations between the IRR, PR, and POR provide further argument in favour of using the PR rather than the POR—that is, consistency of direction of bias relative to the IRR, and the fact that the PR is a conservative measure relative to the IRR.

Estimation of the prevalence ratio

The second focus of the recent debate has been on the appropriate model with which to estimate the PR when adjusting for multiple covariates.[3 6 12 14] Logistic regression is a statistical method which can be used to estimate the probability of disease for a given covariate pattern. The results of logistic regression lead naturally to an estimate of the odds ratio but may also be used to estimate the PR, simply by taking the ratio of the estimated probabilities. This fact has been noted in other publications[6 10] and it has also often been stated[8 11 15] that there is no corresponding variance estimate for the PR. Stromberg[11] states, for instance, that: "As far as I know, there is no useful statistical model for directly estimating a PR with adjustments for several covariates. Such an estimate can be obtained from the logistic model by a straightforward transformation although further research is needed to provide an appropriate confidence interval."

We have derived an expression for the variance of the log of the estimated PR which involves standard results for the variance and

Table 2 Hypothetical cross sectional data

	Disease incidence	Exposed		Disease incidence	Unexposed	
		Diseased	Non-diseased		Diseased	Non-diseased
High prevalence:						
Stratum 1	0.05	44	31	0.01	13	62
Stratum 2	0.09	79	21	0.03	44	56
Stratum 3	0.03	63	87	0.02	48	102
Low prevalence:						
Stratum 1	0.03	31	44	0.006	8	67
Stratum 2	0.015	24	76	0.005	9	91
Stratum 3	0.005	13	137	0.0033	9	141

Data correspond to a 20 year follow up of a cohort which has 75, 100, and 150 exposed and unexposed workers in each stratum with an exodus rate of exposed workers from the cohort of 0.01 and of unexposed of 0.005.

covariance of the logistic regression coefficients. This approach would thus allow the use of widely available logistic regression packages to estimate the PR from cross sectional studies. Appendix 2 gives the derivation of the result. An alternative test based approach to constructing confidence intervals for the PR from logistic regression has been suggested[10] but it is not clear how this could be applied in the presence of effect modification of the odds ratio.

Two other statistical models have been discussed as alternatives to logistic regression. It has been suggested that the PR be estimated (with variance estimates) from proportional hazards regression (the Breslow-Cox model).[3 4 14] This model involves the assumption of Poisson rather than binomial variability but allows the estimation of the PR directly by applying Cox's proportional hazards model to a closed cohort with a constant risk period.[23 24] The method of generalised linear models (with binomial distribution and log link) has also been suggested[14] as a means of modelling the prevalences (and hence the PR). Although this method does have the right distributional assumptions, it may require constrained estimation to avoid prevalence estimates that are >1.[15]

Table 2 contains hypothetical cross sectional data generated by a spreadsheet program with three levels of an effect modifying variable with associated IRRs 5, 3, and 1.5 and overall prevalences of disease of about 45% and 14%. Table 3 shows the different models applied to these data and the resulting effect estimates.

Table 3 Models applied to hypothetical cross sectional data generated by a spreadsheet program

Stratum	High prevalence	Low prevalence
Logistic regression:		
POR (95% CI):		
1	6.8 (3.1 to 14.6)	5.9 (2.4 to 14.3)
2	4.8 (1.3 to 5.5)	3.2 (1.4 to 7.4)
3	1.5 (0.6 to 3.3)	1.5 (0.6 to 3.7)
PR (95% CI):		
1	3.4 (2.0 to 5.8)	3.9 (1.9 to 8.0)
2	1.8 (1.4 to 2.3)	2.7 (1.3 to 5.5)
3	1.3 (0.97 to 1.8)	1.4 (0.6 to 3.3)
Proportional hazards regression:		
PR (95% CI):		
1	3.4 (1.8 to 6.4)	3.9 (1.8 to 8.6)
2	1.8 (1.2 to 2.6)	2.7 (1.2 to 5.8)
3	1.3 (0.9 to 1.9)	1.4 (0.6 to 3.4)
Generalised linear model (with a log link and binomial distribution):		
PR (95% CI):		
1	3.4 (2.0 to 5.8)	3.9 (1.9 to 8.0)
2	1.8 (1.4 to 2.3)	2.7 (1.3 to 5.5)
3	1.3 (0.97 to 1.8)	1.4 (0.6 to 3.3)

The high prevalences near 50% are plausible for, say, musculoskeletal and many other health outcomes measured in cross sectional studies. The low prevalence is near the typical cut off (10%) applied in practice to satisfy the rare disease assumption. Yet, the differences between the two effect measures are still quite large as the point estimate moves away from the null value.

It should be noted that the above results represent saturated statistical models which are equivalent to stratum specific 2×2 analyses. The advantage of a model representation is that it may be possible to represent the data more parsimoniously and, in fact, all models suggest pooling the exposure coefficients for the first two strata. For the PH and GLM methodologies considered above this would result in a pooled PR estimate for strata 1 and 2 and for logistic regression a pooled POR estimate over these strata, but nevertheless different stratum-specific PRs. This point is considered again later.

Limitations to the use of logistic regression for modelling the PR

Although the logistic model does provide a means of obtaining point estimates of prevalence and the prevalence ratios on a stratum specific basis, there is no natural way to obtain a pooled or adjusted effect estimate for either point or interval estimates of the PR in situations where the stratifying variable is related to outcome. So it is necessary to elaborate the analysis for each stratum resulting in the equivalent of an epidemiological effect modification analysis. This can be onerous in practice if there are many strata and covariates. An extreme instance of this problem would arise in the presence of a continuous covariate, X, in which case there would potentially be a different estimate of the PR (and its variance) for each level of X.

The data in table 4 provide an example of such a situation. Again, we have a setting with high prevalence (about 47%), with crude POR = 4.4 and PR=2.0. Table 5 shows the resulting logistic regression analysis and the resulting estimates.

The fitted model for the logit of prevalence of disease is:

$$\text{Logit}(p)= -0.4055+1.7918E-0.4418S-0.539\ S{*}E$$

where E=1 for exposed and 0 for unexposed, S=0 for stratum 1 and 1 for stratum 2.

All four coefficients in the logistic regression model are significant at the 1% level. This is

Table 4 Further hypothetical cross sectional data

	Exposed		Unexposed	
	Diseased	Non-diseased	Diseased	Non-diseased
Stratum 1	400	100	320	480
Stratum 2	300	200	300	700

Table 5 Logistic regression estimates

Stratum	PR (95% CI)	POR (95% CI)
1	2.0 (1.8 to 2.2)	6.0 (3.2 to 14.4)
2	2.0 (1.8 to 2.3)	3.5 (2.8 to 4.4)

necessitated by the presence of effect modification in the odds ratio. There is no natural way of collapsing the model to recognise the fact that there is neither confounding nor effect modification of the PR.

Both proportional hazards regression and generalised linear models applied to these data would result in the fitting of a single common PR to the strata, respectively:

PR=2.0, 95% CI 1.8 to 2.2; PR=2.0, 95% CI 1.9 to 2.2.

This example illustrates a point that we have already raised—namely, that both choice of effect measure and of statistical model can have an impact on whether a covariate is identified as a confounder or effect modifier.

We have carried out all the statistical analyses with procedures from the SAS statistical package.[23 24] Specifically: PROC LOGISTIC, PROC PHREG, and PROC GENMOD.

Summary

As noted in the introduction, cross sectional studies are sometimes used for descriptive purposes, when the prevalence is clearly the appropriate measure of disease frequency, and no link to incidence is sought. In these situations, the appropriate ratio measure is the PR, and these results show that the POR will not provide a consistent approximation to the PR. Hence, the POR should not be used in these settings.

In aetiological research settings, consideration of other publications and our results leads us to conclude that selecting the PR as an effect measure of choice over the POR is to be recommended. It is more interpretable and more consistent for estimating the true effect, taken as the IRR. The POR is difficult to interpret as an effect measure (outside of certain specific settings). Furthermore, the POR is inconsistent in its relation with the IRR, sometimes overestimating and sometimes underestimating. Finally, use of the POR will not necessarily lead to the same conclusions as from the PR about effect modification or confounding.

On the other hand, the more appropriate and consistently behaved PR is more difficult to estimate in the multivariate setting. The possible statistical methods for multivariate analysis of the PR as a measure of the effect of an exposure upon disease with several covariates in cross sectional studies are:

(1) Logistic regression, with estimation of the PR from the estimated prevalences, and the variance of the PR from the method we propose. Logistic regression has the advantage of familiarity and wide availability, but may be unwieldy in that, if there is confounding or effect modification in the POR, or even a significant independent covariate effect, effect measures based on the PR must be estimated separately for each stratum (whether or not there is effect modification in the PR) and perhaps pooled on an ad hoc basis. A weighted average of the PRs across the strata could, for instance, be computed, with the weights being proportional to the inverse of the estimated variance of each PR. This is essentially a directly pooled estimate,[19] with the weights being based on a statistical model. The well known Mantel-Haenszel estimate is also a weighted average, except there the weights are based on the individual cell counts rather than fitted values from a model (and so in statistical terms are based on the saturated model).

(2) Proportional hazards regression to directly estimate the PR, but with wider or less precise interval estimates.

(3) A model for the individual (log) prevalences based on the binomial distribution. Generalised linear models provide a way of doing this, but with the problem that the prevalences may have to be artificially constrained to be between 0 and 1 (so a potentially more complicated procedure and one that is not necessarily maximum likelihood). It should further be noted that the SAS software that we used to apply generalised linear models is only applicable to grouped data.

There is clearly still room for exploration of the three statistical methods described and perhaps development of alternatives. Logistic regression has the virtue of already being a familiar tool for many epidemiologists. With the results that we have provided above, it will be possible to estimate the PR (both point and interval estimates) as well as the POR by this method. However, if there are many covariate strata, the results for estimating the PR by logistic regression may be quite cumbersome. In this case, we recommend either a generalised linear model or proportional hazards regression; each has its strengths and limitations.

Appendix

Appendix 1: Derivation of the association between the IRR, PR and POR

Let a_t and b_t represent respectively the numbers in the exposed group of disease free and diseased workers in the cohort at time t and let c_t be the number of exposed diseased people who have left the cohort by this time. Similarly, d_t, e_t, f_t represent the equivalent quantities for the unexposed people.

At the start of the follow up we have a_0=n, d_0=m and b_0=c_0=e_0=f_0=0. The system of differential equations that governs the behaviour of a_t, b_t, c_t is given by:

$$da_t/dt = \alpha_1\ b_t - p_1 a_t$$

$$db_t/dt = p_1\ a_t - \alpha_1 b_t$$

$$dc_t/dt = \alpha_1 b_t$$

with similar expressions for d_t, e_t, f_t.

Solving the system of differential equations with the given starting conditions yields:

$$a_t = \frac{n}{(\alpha_1 + p_1)}(\alpha_1 + p_1 e^{-(\alpha_1+p_1)t})$$

$$b_t = \frac{np_1}{(\alpha_1 + p_1)}(1 - e^{-(\alpha_1+p_1)t}) \qquad (1)$$

$$c_t = \frac{n\alpha_1 p_1}{(\alpha_1 + p_1)}(t - \frac{1 - e^{-(\alpha_1+p_1)t}}{(\alpha_1 + p_1)})$$

with again similar expressions for d_t, e_t, f_t.

Using this notation, we have:

$$PR_t = \frac{\dfrac{b_t}{a_t + b_t}}{\dfrac{e_t}{d_t + e_t}} = \frac{mb_t}{ne_t}$$

$$POR_t = \frac{b_t d_t}{a_t e_t} \qquad (2)$$

$$IRR_t = \frac{\dfrac{b_t + c_t}{\int_0^t a_u du}}{\dfrac{e_t + f_t}{\int_0^t d_u du}} = \frac{p_1}{p_2}$$

A comparison of these expressions, substituting (1) into (2) yields:

$$(i) PR_t = IRR \; \frac{(a_2 + p_2)\,(1 - e^{-(a_1+p_1)t})}{(a_1 + p_1)\,(1 - e^{-(a_2+p_2)t})}$$

$$(ii) POR_t = PR_t \; \frac{(a_1 + p_1)\,(a_2 + p_2\, e^{-(a_2+p_2)t})}{(a_2 + p_2)\,(a_1 + p_1\, e^{-(a_1+p_1)t})}$$

$$(iii) POR_t = IRR \; \frac{(a_2 + p_2\, e^{-(a_2+p_2)t})\,(1 - e^{-(a_1+p_1)t})}{(a_1 + p_1\, e^{-(a_1+p_1)t})\,(1 - e^{-(a_2+p_2)t})}$$

Appendix 2: Derivation of an expression for the variance of the estimated log PR

Let x denote the covariate pattern in the numerator of the PR and y the covariate pattern in the denominator. In the simplest case of a single dichotomous exposure variable and no covariates, we would have x = (1,1) and y=(1,0), with the first component in each vector denoting the intercept and the second the presence (absence) of the exposure. Let p_1 denote the probability of disease under covariate pattern x and p_2 that for covariate pattern y. Logistic regression assumes the following form for p_1 (and, equivalently, p_2):

$p_1 = e^{x\beta}/(1 + e^{x\beta})$.

A standard Taylor series development yields the following expression for the variance of the log of the estimated PR: $\sum_i \sum_j (x_i(1-p_1)-y_i(1-p_2))(x_j(1-p_1)-y_j(1-p_2))\mathrm{cov}(b_i,b_j)$ where b_i is the estimated logistic regression coefficient corresponding to the i'th element of x and y and $\mathrm{cov}(b_i,b_j)$ is the covariance between the i'th and j'th estimated coefficients.

Now, the equivalent expression for the variance of the estimated log odds ratio from logistic regression is:

$\sum_i \sum_j (x_i-y_i)(x_j-y_j)\mathrm{cov}(b_i,b_j)$

from which it easily follows that var(log(POR))>var(log(PR)).

1 Keiding N. Age-specific incidence and prevalence: a statistical perspective. *J R Statist Soc A* 1991;**154**:371–412.
2 Alho JM. On prevalence, incidence and duration in general stable populations. *Biometrics* 1992;**48**:587–92.
3 Lee J, Chia KS. Estimation of prevalence rate ratios for cross-sectional data: an example in occupational epidemiology. *Br J Ind Med* 1993;**50**:861–2.
4 Lee J. Odds ratio or relative risk for cross-sectional data. *Int J Epidemiol* 1994;**23**:201–3.
5 Axelson O. Some recent developments in occupational epidemiology. *Scand J Work Environ Health* 1994;**20**:9–18.
6 Stromberg U. Prevalence odds ratio v prevalence ratio. *Occup Environ Med* 1994;**51**:143–4.
7 Axelson O, Fredriksson M, Ekberg K. Use of the prevalence ratio v the prevalence odds ratio as a measure of risk in cross sectional studies. *Occup Environ Med* 1994;**51**:574.
8 Lee J, Chia KS. Use of the prevalence ratio *v* the prevalence odds ratio as a measure of risk in cross sectional studies. *Occup Environ Med* 1994;**51**:841.
9 Hughes K. Odds ratios in cross-sectional studies. *Int J Epidemiol* 1995;**24**:463–4.
10 Osborn J, Cattaruzza MS. Odds ratio and relative risk for cross-sectional data. *Int J Epidemiol* 1995;**24**:464–5.
11 Stromberg U. Prevalence odds ratio v prevalence ratio - some further comments. *Occup Environ Med* 1995;**42**:143.
12 Axelson O, Fredriksson M, Ekberg K. Use of the prevalence ratio v the prevalence odds ratio in view of confounding in cross sectional studies. *Occup Environ Med* 1995;**52**:494–6.
13 Lee J, Chia KS. Prevalence odds ratio v prevalence ratio—a response. *Occup Environ Med* 1995;**52**:781–4.
14 Zocchetti C, Consonni D, Bertazzi PA. Estimation of prevalence rate ratios from cross-sectional data. *Int J Epidemiol* 1995;**24**:1064–5.
15 Lee J. Estimation of prevalence rate ratios from cross sectional data: a reply. *Int J Epidemiol* 1995;**24**:1066–7.
16 Nurminen M To use or not to use the odds ratio in epidemiologic analysis? *Eur J Epidemiol* 1995;**11**:365–71.
17 Zocchetti C, Consonni D, Bertazzi PA. Relationship between prevalence rate ratios and odds ratios in cross-sectional studies. *Int J Epidemiol* 1997;**26**:220–3.
18 Greenland S. Interpretation and choice of effect measures in epidemiologic analysis. *Am J Epidemiol* 1987;**125**:761–8.
19 Rothman KJ. *Modern epidemiology*. Boston: Little, Brown, 1986.
20 Eisen E. Healthy worker effect in morbidity studies. *Med Lav* 1995;**86**:125–38.
21 Steineck G, Ahlbom A. A definition of bias founded on the concept of the study base. *Epidemiology* 1992;**3**:477–82.
22 Nurminen M. On the epidemiologic notion of confounding and confounder identification. *Scand J Work Environ Health* 1997;**23**:64–71.
23 Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B* 1972;**34**: 187–220.
24 Breslow NE. Covariance analysis of censored survival data. *Biometrics* 1974;**30**:89–99.
25 SAS Institute. *SAS/STAT User's guide, version 6.09, 4th ed.* Cary, NC: SAS Institute, 1989.
26 SAS Institute. *The GENMOD procedure. Release 6.09.*Cary NC: SAS Institute, 1993. (SAS technical report P-243.)